AD NUMBER

AD253952

LIMITATION CHANGES

TO:

Approved for public release; distribution is unlimited.

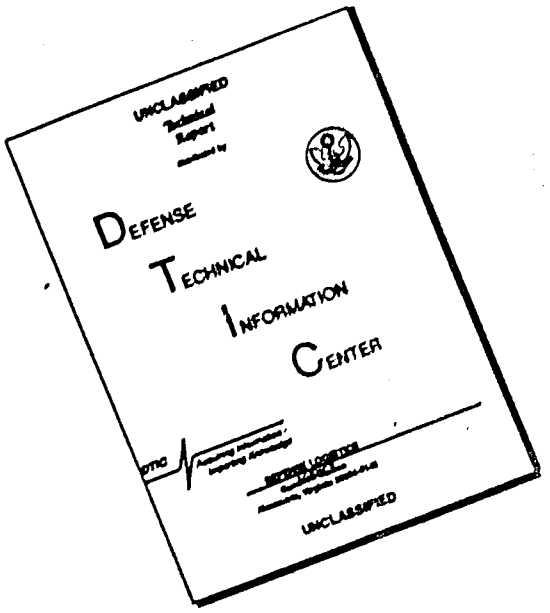
FROM:

**Distribution authorized to U.S. Gov't. agencies and their contractors;
Administrative/Operational Use; DEC 1960. Other requests shall be referred to Office of Naval Research, Arlington, VA 222203.**

AUTHORITY

ONR ltr 7 Sep 1970

DISCLAIMER NOTICE



THIS DOCUMENT IS BEST QUALITY AVAILABLE. THE COPY FURNISHED TO DTIC CONTAINED A SIGNIFICANT NUMBER OF PAGES WHICH DO NOT REPRODUCE LEGIBLY.

EXPERIMENTS ON DECISION MAKING AND OTHER STUDIES

Edited by
Donald W. Taylor

Technical Report 6

Prepared under Contract Nonr 609(20)
(NR 150-166)
for
Office of Naval Research

DEPARTMENT OF INDUSTRIAL ADMINISTRATION
AND DEPARTMENT OF PSYCHOLOGY
YALE UNIVERSITY
NEW HAVEN, CONN.

December, 1960

EXPERIMENTS ON DECISION MAKING

AND OTHER STUDIES

Edited by

Donald W. Taylor

Technical Report 6

Prepared under Contract Nonr 609(20)
(NR 150-166)
for
Office of Naval Research

DEPARTMENT OF INDUSTRIAL ADMINISTRATION
AND DEPARTMENT OF PSYCHOLOGY
YALE UNIVERSITY
NEW HAVEN, CONN.

December, 1960

FOREWARD

The studies reported here were carried out under Project NR 150-166 and supported in whole or in part by Contract Nonr 609(20) between Yale University and the Office of Naval Research. Appreciation for their work in connection with the contract is extended to Dr. Glenn L. Bryan, Head, and to Dr. John Nagay, Assistant Head, Personnel and Training Branch, Psychological Sciences Division.

Donald W. Taylor
Professor of Psychology

TABLE OF CONTENTS

AMOUNT AND GENERALITY OF INFORMATION-SEEKING BEHAVIOR IN SEQUENTIAL DECISION MAKING AS DEPENDENT ON LEVEL OF INCENTIVE

Donald R. Worley

Empirical research in the field of decision making has dealt primarily with simple choice situations. In such situations, a subject is given a certain amount of information and then is asked to make a choice between alternatives A and B. The sequential decision situation, however, is more complex. In this situation the subject not only attempts to choose which alternative is correct but also decides when he has sufficient information to make a choice. He is allowed to seek information before making a decision. Thus at any point in the decision process he may choose A or B or choose to seek more information.

Many real-life situations involve sequential decision making. For instance, a general on the battle field must choose between two or more possible maneuvers. If the intelligence reports have not all been received, the general must decide which maneuver to employ on the information at hand or wait until more information is received. If he decides to wait, his decision may be more correct, but if he makes a decision based on the information at hand, he may gain the advantage of better timing. At some point also the general must conclude that the value of the next piece of information is not great enough to require that he wait for it before making his decision.

Most studies of sequential decision making have been analyses by statisticians. An empirical study, however, was carried out by Pruitt (1957). Pruitt's experiment included two types of sequential decision-making tasks. In the first the subject took his information from a machine which flashed lights randomly either in the proportion 60 per cent red and 40 per cent green or 40 per cent red and 60 per cent green. Every time the subject pressed a button either a red or a green light would go on. Pruitt employed two variations of this task, but in both variations the subject was required to decide which of the settings had appeared. The other type of task was judging line length. Two lines were presented on each of 20 slides. On all the slides, the line on one side, either the right or the left, was consistently the longer. The subject was to decide which line was longer after looking at any number of slides up to 20. In some of the task repetitions, the subject was scored according to a point system.

The less information he required before making a decision, the more points he scored. Information in the machine task was defined as the difference between the number of red and green lights. Information in the line judging task was defined as the number of slides seen before making a decision.

The results showed that there was a significant correlation between all the task variations in the amount of information sought, and that the effect of the incentive was to reduce the amount of information which the subjects required before making their decisions. It is of interest to note that the correlations between the two machine tasks (both incentive tasks), and the incentive line judging task were higher than the correlations between the former and the no-incentive line judging tasks; however, Pruitt did not discuss this difference and it was not statistically significant.

The purpose of the present experiment was to explore further both the generality of information-seeking behavior and the effects upon such behavior of level of incentive. Money was employed as the incentive in the thought that its importance for motivation might be greater than the award of points employed by Pruitt. In view of Pruitt's findings, the following predictions were made:

(1) A high level of incentive will cause the subjects to seek less information than will a low level of incentive.

(2) Amount of information sought on one type of sequential task will correlate with that sought on other types of such tasks.

(3) The correlation of information sought in one task with information sought in another task will be higher under high incentive conditions than under low incentive conditions.

Procedure

An experiment was devised to test the above predictions. Seventy-two subjects were used; all were undergraduates of Yale University. Every subject was given three types of decision making tasks and each task was repeated three times. The tasks in the order given are described below.

In the Dice Task the subject was presented with three dice. He was told that one of the dice was "crooked" and that it favored sixes and ones. The data show that sixes and ones came up 39 per cent of the time, so the bias was slight and somewhat less than the experimenter had hoped for. The subject was asked to decide which die was "crooked" by throwing the dice and recording the outcomes

on an answer sheet. One trial consisted of three throws, one of each die, and the subject was limited to 25 trials.

In the Marbles Task the subject was presented with three urns of marbles. Two of the urns each contained 50 red and 50 yellow marbles, and the subject was informed of these proportions. In the third urn there were 55 red and 45 yellow marbles. The subject was told only that there were more red than yellow marbles in one of the urns. The task was to decide which of the urns had more red marbles by drawing the marbles one at a time out of the urns. After drawing a marble the subject had to replace it in the urn. Thus the above proportions remained constant throughout the task. One trial consisted of three draws, one from each urn, and the subject was limited to 25 trials.

In the Clues Task the subject was given a stack of note cards. On each card was written a clue which would help to identify a specific object. The subject was required to look at the clues in a given order until he had seen enough clues to identify the object referred to. A trial was considered as one clue. Tables 1, 2, and 3 present a list of the clues in the order given and the correct answer for each of the three tasks respectively.[1]

The subjects were divided into high and low incentive groups. The high incentive group could earn a maximum of 50 cents per task repetition. This maximum was earned if the subject made a correct decision after one trial. The amount was reduced by two cents per trial thereafter. The low incentive group could earn a maximum of 10 cents per task repetition, and this maximum could be earned if the subject made the correct decision after any of the first five trials. After trials 6-10 inclusive he could earn eight cents; after trials 11-15 he could earn six cents, and so on.

The subjects were informed of the maximum amount which they could earn for each repetition and were told that the more trials they took before making a decision the less money they would receive. They were not told of the exact nature of the incentive function. There were two reasons for keeping this latter information from the subject. First of all, it was thought that a sophisticated subject might be able to calculate approximately the most efficient strategy for each task. Since he did not know the probable outcomes in the marbles and dice tasks and did not know the incentive function, such a strategy would be impossible

---

[1]The author wishes to acknowledge the contribution of John S. Roberts, Jr. in developing two of the Clues Tasks employed. Since, however, the three tasks actually used differed somewhat from those employed by Roberts (1960), they are reproduced here in full.

Table 1

First Clues Task

Clues

1. Vegetable
2. Non-living
3. Manufactured
4. Can be held in the hand
5. Used by both sexes
6. Used by all age groups
7. Produced in nearly all countries of the world
8. Mainly a wood product
9. Printed matter
10. Made in a variety of designs
11. Comes in a variety of colors
12. Weighs less than a pound
13. Essential in the performance of a specified service
14. A number of the same type are made together
15. Not any sort of ticket
16. Comes in different denominations
17. In U.S. is made under government supervision
18. Has monetary value
19. Many types commemorate some person
20. Not any form of currency
21. Many types commemorate an event
22. Comes in various sizes
23. Usually every dimension is less than two inches
24. Has glue on unprinted side
25. Its edges are perforated
26. Sold by the post office

ANSWER: POSTAGE STAMP

to find. The experiment was intended to study information-seeking behavior rather than the degree to which individual strategies conformed to a mathematical model based on expected value. Thus the experimenter was interested in how subjects perceived information and how they acted upon perceived information, rather than how they used objective information in comparison with how they should have used objective information. This is also the reason why it was not necessary for the experimenter to control the outcomes, in drawing of the marbles or rolling of the dice, even though Pruitt's findings involve tasks where the outcomes were controlled. It is evident that on any given trial many outcomes are possible and the possible outcomes have various probabilities of indicating the correct decision. By keeping the knowledge of the probable outcomes and the incentive functions from the subjects the emphasis was placed on perceived rather than objective information.

Table 2

Second Clues Task

Clues

1. Mineral
2. Manufactured
3. Used by humans
4. Used more by adults
5. Used by both sexes
6. Can be used day or night
7. Can be used any time of year
8. Can be used in the home
9. Can be used at work
10. Can be held in the hand
11. Made partly of metal
12. Can be dangerous
13. Sometimes used for pleasure
14. Sometimes use of it can be illegal
15. Comes in a variety of sizes
16. Pointed
17. Has a moving part
18. Not a pair of scissors
19. Can be used on animals
20. Can be used to save a life
21. Not any kind of knife or scalpel
22. Can cause pain
23. Sight of it can be frightening
24. Used in hospitals
25. Used by dope addicts
26. Holds medicine

ANSWER: HYPODERMIC NEEDLE

The second reason for not telling the subject the exact nature of the incentive function was to keep the subject from making some decision beforehand concerning the amount of money which he would regard as fair pay for his efforts. If he did this he would be likely to limit the number of trials he took according to his notion of fair pay. Also, if the subject knew what his probable payment was on one task repetition, this knowledge might affect his other decisions. Thus he might vary the number of trials he took, consciously attempting to make the number of trial average at a level which would earn what he thought was fair pay. For this same reason the subjects were not informed of the correctness of their decisions or the amounts which they had earned until the experimental session was over.

Table 3

Third Clues Task

Clues

1. Vegetable
2. Non-living
3. An object
4. Manufactured
5. Always made in the same basic shape
6. No moving parts
7. Can be used by both sexes
8. Used for recreational purposes
9. Held while being used
10. Used in a particular sport
11. Comes in contact with another object when being used
12. Is not thrown
13. Does not roll
14. Not any sort of ball
15. Sport in which it is used is played outdoors
16. Longer than it is wide
17. Made of wood
18. Made in standard sizes
19. Not any sort of racquet
20. Made from a single piece of material
21. Not a hockey stick
22. Round in one dimension
23. Tapered
24. Made by being turned on a lathe
25. Sport in which it is used played in spring and summer
26. Sport in which it is used recently added a third league

ANSWER: BASEBALL BAT

The subjects of both incentive groups were told that they would be given only the amount which they earned for participating in the experiment. It should be noted that Pruitt paid his subjects a flat rate for participating and no money was promised for earning points in his tasks. The subjects in this study were paid what they earned or paid a dollar, whichever was the larger amount, but at the beginning they were promised only what they could earn. This was meant to enhance the effects of the money incentive.

Results

One measure of information-seeking behavior is the number trials a subject takes before making a decision. The mean number of trials taken in each task repetition for each incentive group is presented in Table 4.

Table 4

Mean Number of Trials Taken by a Subject before Making a Decision

		High Incentive	Low Incentive
Dice	1	7.94	6.33
	2	7.33	6.66
	3	8.14	7.31
	Mean	7.81	6.76
Marbles	1	12.83	11.08
	2	12.78	11.61
	3	12.47	11.39
	Mean	12.70	11.36
Clues	1	14.70	14.03
	2	18.08	17.64
	3	16.22	15.22
	Mean	16.33	15.63

The *t*-test was applied to the differences between the means of three repetitions of a task in the high and low incentive groups, but none were statistically significant. If the means of successive repetitions of each task were independent, the fact that the differences between the nine pairs of means were in the same direction would be significant at the .002 level. Since the means intercorrelate, the sign test is not applicable.

To test the reliability of the number of trials as a measure, the results of each of the three repetitions of each type of task were intercorrelated. The correlations were then transformed to *z* scores. The mean of the *z* scores was then transformed again to *r* to give the mean reliability. Since the mean *r* was the reliability of one task repetition, the Spearman-Brown formula was applied to give an estimate of the reliability of a mean based on three task repetitions. The estimated reliabilities for each task and each incentive group are presented in Table 5.

Table 5

Reliability of Number of Trials Taken
As a Measure of Information-Seeking Behavior

	High Incentive	Low Incentive
Dice	.77	.68
Marbles	.87	.84
Clues	.78	.71

The correlations between the types of tasks of the number of trials taken produced the results shown in Table 6. Two of the correlations were significant at the .01 level and two at the .05 level. The other two correlations were not significant but were in the direction predicted.

Table 6

Correlation between Tasks in Information Sought
As Measured by Number of Trials Taken

	High Incentive	Low Incentive
Dice-Marbles	.63**	.38*
Marbles-Clues	.28	.42**
Dice-Clues	.23	.35*

* Significant at the .05 level
** Significant at the .01 level

The difference in the Dice-Marbles correlations between the high incentive and low incentive groups was in the direction predicted but reached only the .16 level of significance. The differences between the other two pairs of correlations was in the opposite direction from that predicted and was not significant.

Another measure of information-seeking behavior in the Dice and Marbles Tasks is the difference between the two largest outcomes that a subject obtained before making a decision in a task repetition. Thus, for example, on the twentieth trial in the Marbles Task, the subject to that point may have obtained 9 red (and 11 yellow) marbles from the first urn, 11 red from the second, and 13 red from the third urn; the difference between the two largest outcomes at this point

would be 13 minus 11, or 2. The estimated reliabilities of this measure are presented in Table 7.

Table 7

Reliability of Difference in Outcomes
As a Measure of Information-Seeking Behavior

	High Incentive	Low Incentive
Dice	.60	.50
Marbles	.00	.48

The correlations between the types of tasks, using the differences between outcomes as a measure for the Dice and Marbles Tasks and trials as the measure for the clues tasks, are shown in Table 8. The difference in the Dice-Marbles correlations between the high incentive and low incentive groups was in the direction predicted and was statistically significant. The differences between the other two pairs of correlations were not significant.

Table 8

Correlation Between Tasks in Information Sought
As Measured by Differences in Outcomes

	High Incentive	Low incentive
Dice-Marbles	.45**	.00
Marbles-Clues	.23	.28
Dice-Clues	.26	.22

** Significant at the .01 level

Discussion

It was noted that the mean number of trials in all repetitions of the tasks was larger for the high incentive group. Although none of the differences were significant, the direction of the differences suggest that the effect of high incentive is to cause the subject to seek more information. This effect is contrary to the first prediction and is in the opposite direction from that which Pruitt (1957) reported. He, however, did not analyze the differences between

the means of the incentive and no-incentive conditions. Furthermore, his study employed a points rather than a money incentive. The differences in mean number of trials between the three types of tasks seem to be due to the nature of the tasks themselves.

The high reliability of number of trials as a measure (Table 5) in comparison to the reliability of difference between outcomes (Table 7) indicates that the former is a better measure of information-seeking behavior. Also the high reliability suggests that the subjects perceived information in terms of number of trials even though the objective information in the Dice and Marbles Tasks was the actual outcome which was obtained in each repetition.

With two exceptions, the correlations between the types of tasks were statistically significant (Table 6), thus confirming the second prediction. This result implies that information-seeking behavior generalizes between sequential decision-making situations. Thus a person who seeks a large amount of information in one situation will seek a relatively large amount in a similar situation. Even the clues task, which differs from the other two in that it tests the ability to use verbal information rather than sampling information, correlated significantly in two cases. The correlations which were not significant were nearly so; one was significant at the .10 level whereas the other just failed to reach that level.

The difference between the Dice-Marbles correlations in the high and low incentive groups (Table 6), although only significant at the .16 level, does not allow us to reject the prediction that high incentive increases the generality of information-seeking behavior between similar tasks. With the difference in outcome as the measure, the difference between the high and low incentive Dice-Marbles correlations (Table 8) is also in the same direction and is statistically significant. However, in view of the relatively low reliability of the latter measure, it is questionable that the prediction has been confirmed. Moreover, the difference between the high and low incentive groups in the Marbles-Clues and Dice-Clues correlations contradict the prediction that high incentive will increase the generality of information-seeking between tasks.

Summary

An experiment was designed to test the effects of high and low incentive on information-seeking behavior in sequential decision-making situations. Three

types of decision-making tasks were used and it was found that the number of trials taken before making a decision is a reliable measure of information-seeking behavior. The following results were obtained:

(1) The first prediction was not confirmed. Although the differences were not significant, the mean number of trials for each of the three repetitions of all three tasks was larger with high than with low incentive, thus suggesting that increasing incentive may increase rather than, as Pruitt had found, reduce amount of information sought.

(2) The amount of information sought, measured as number of trials taken, correlated significantly between the different types of task in four cases out of six and were in the direction predicted. The other two correlations, although not significant were also in the direction predicted. Thus, the second prediction was confirmed.

(3) The third prediction that the correlations between tasks would be larger with high than with low incentive was not confirmed.

References

Pruitt, D. An exploratory study of individual differences in sequential decision making. Ph.D. Dissertation. Yale University, 1957.

Roberts, J.S., Jr. Information seeking in sequential decision making as dependent upon test anxiety and upon prior success or failure in problem solving. See pp. 26-36 of this report.

MAXIMIZATION OF UTILITY IN ECONOMIC DECISIONS UNDER RISK[1]

Earl B. Hunt

Economic decisions may be viewed as attempts to maximize utility. In the risk situation the subject must choose between several alternatives, each with a probabilistically determined set of outcomes. A payoff, in terms of a specified commodity, is associated with each outcome. In this situation a subject should choose alternatives which maximize the expected return in utility units. No prediction of behavior can be made unless the relation between utility and the commodity used for payment is known.

Utility is a psychological concept, analogous to sensory scaling. The function relating utility to money is of particular interest in economic decision mainly because of the central role of money as an index for other commodities. In 1758 Daniel Bernoulli argued that the "simplest" assumption, that utility is a linear function of money, would lead to the untenable St. Petersburg paradox. Bernoulli suggested that there was a logarithmic relation between utility and money. Later writers have generally agreed that this is reasonable, although Freidman and Savage (1948) pointed to some situations for which a doubly inflected utility curve seemed reasonable. Edwards (1954) noted that most economists' discussions of utility were examples of "armchair theorizing." They considered important behavior (e.g. stock investment, insurance buying) but were not supported by experimental evidence. Recently several experimental attempts to determine and make use of utility functions have been reported (Mosteller and Nogee, 1951; Davidson, Suppes and Siegel, 1957; Suppes and Walsh, 1959). Through sophisticated scale construction techniques, individual curves for the relation between utility and money have been obtained. These studies have involved data from real economic decisions and, unavoidably, were limited by the amount of money that could be used. Subjects' wins and losses were generally in the range of two or three dollars or less, sometimes only pennies. In addition the subjects were drawn from available, rather than purposely selected, sources. In summary, the implications of

---

[1]This research was performed while the author held a General Electric Company or a National Science Foundation predoctoral fellowship. The research was partially supported by Nonr 609(20) between Yale University and the Office of Naval Research.

the concept of utility have been examined by discussion of important economic decisions or by experimental study of minor ones.

The typical utility investigation has been concerned with the utility function of an individual. In many cases behavioral scientists wish to make statements about averages. Group trends could be examined to see whether or not they agreed with predictions generated from a particular (utility) model describing the individuals in the group. Individual behavior could also be examined to determine the extent to which the average trend represented the behavior of specific individuals. Even if the model should fail at this point (i.e. if individual prediction was not successful) it would be of use in generating predictions about the group.

The study reported here is an attempt to gain experimental evidence about economic decisions in a situation intermediate between "armchair discussion" and real, small wagers. Students in an advanced undergraduate economics class were asked to "role play" the part of investment counselors advising a hypothetical firm on the investment of surplus capital. By this method subjects offered opinions on non-trivial economic decisions within an experimental setting. Both the setting and subjects were chosen to increase the face validity of role playing results. The economics students were familiar with the experimental task and accepted as legitimate an inquiry into economic behavior. It is not inconceivable that some of the subjects in this experiment will shortly play a real life role similar to the one required of them in a research setting!

Description of the Task[2]

The subjects' task was to recommend distribution of sums from a fixed total in the purchase of one or more of sixteen possible bond issues. The issues varied in amount of interest paid and probability of payment. These were covaried so that all issues had the same expected monetary return. The situation was an example of choice between alternatives varying in risk and possible profits but equal in terms of expected profit.

---

[2]An analysis of the experimental task in terms of maximization of concave, convex, and linear utility functions was suggested by Jacob Marschak. Alan Manne suggested the applicability of Markowitz's analysis of utility maximization. This assistance is greatly appreciated. Neither Professor Marschak nor Professor Manne has approved the final draft of this paper and they do not necessarily agree with the analysis or the conclusions reached. For these the author bears the sole responsibility.

The implications of three possible forms of a utility curve; convex, linear, and concave, were considered as they applied to the experimental alternatives. (Cases in which a doubly inflected utility curve appears reasonable [cf. Friedman and Savage, 1948] did not occur in the experiment.) The three possible forms were defined as follows:

(1) A utility function, $u(x)$, is convex for a commodity, x, if, for any set of values of the commodity x_i, and any set of real, non-negative numbers a_i such that the $\Sigma a_i = 1.00$,

$$1. \quad \Sigma a_i \cdot u(x_i) > u(\Sigma a_i x_i).$$

(2) The utility function is linear if Equation 1 is an equality.

(3) The utility function is concave if the inequality of Equation 1 is reversed.

Let p_i be the probability that the ith investment alternative will pay. Since the expected value of all investments is the same, it can be set equal to 1.00 by changing the scale used to describe profits. If x_i commodity units are invested in the ith alternative, the possible profits for this investment, on the new profit scale, are (x_i/p_i) with probability p_i or 0 with probability $1-p_i$. However the investor has the option of distributing his capital over k possible investments. Define a set of subscript vectors of k elements whose entries are zeroes or ones. Let p_{Si} denote the probability of simultaneous payment for the alternatives whose corresponding entries in the subscript vector S_i are one and simultaneous nonpayment for all other alternatives. [For example, if $k = 2$, $S_1 = (1,1)$, the p_{Si} is the probability that alternatives 1 and 2 pay simultaneously.] If M is defined as a vector with elements (x_i/p_i) for all i, $i = 1.....k$, then the probability distribution for payoffs associated with any investment decision M, which specifies the x_i) is given by the set of C terms

$$2. \quad \{S_i' M \text{ with probability } p_{Si}\}, \quad i = 1.....C$$

Since a constant sum of money is to be invested let $\Sigma x_i = 1.00$. An investment decision, M, is diversified if there is no $x_i = 1.00$. The utility of an investment decision associated with a particular outcome is $u(S_j' M)$. Note that $(S_j' M)$ is a scalar.

The expected utility of a particular decision, M, is

$$3. \quad EU_M = u\left(\sum_j^C p_{Sj} \, [S_j' M]\right).$$

If the decision is to diversify, Equation 3 must be such that, for this $\underline{M}$, no $\underline{x}_i = 1.00$ and

$$4. \quad EU_M \geq \max \left\{ p_1 \, u(1/p_1) \ldots\ldots\ldots, p_k \, u(1/p_k) \right\}.$$

The terms on the right of Equation 4 represent the utilities of all possible non-diversified decisions.

Suppose $\underline{u}(\underline{x})$ is convex. If $\underline{p}_i \neq \underline{p}_j$

$$5. \quad p_i \cdot u(1/p_i) \neq p_j \cdot u(1/p_j).$$

If $\underline{u}(\underline{x})$ is convex or linear and a diversified decision is made

$$6. \quad EU_M = u\left(\sum_i^k x_i p_i \cdot (1/p_i) \right)$$

represents the utility of the expected value of this decision. From the definition of convex and linear functions

$$7. \quad EU_m \leq \sum_i^k x_i \cdot p_i \, u(1/p_i) \, .$$

But the right hand term of Equation 7 is a weighted average of the terms on the right hand side of Equation 4 and must be less than some one of them under the restraint of Equation 5. Since Equation 5 holds for any decision other than diversification over identical alternatives, convex utility functions do not permit diversification. If the utility function is linear, 7 and 5 are equalities and the decision maker cannot discriminate between any possible $\underline{M}$ vectors representing decisions.

In order to maximize a concave utility function an investor must minimize the variance of the probability distribution of outcomes. The argument which will be used to prove this is adapted from Markowitz (1959, esp. pp. 286-288).

Let $\underline{R}$ represent the expected mean return of an investment decision (i.e. a distribution of money over a set of independent alternatives) and $\underline{R}^2$ represent the expected sum of squares of the components of the return. The utility of the distribution of possible outcomes may be approximated by the quadratic function[3]

$$8. \quad u(R) = c + a\,R - bR^2 \quad (a,b, \quad \text{positive numbers})$$

in which $\underline{c}$ represents the point of zero utility. At any given level of $\underline{R}$, maximization of $\underline{u}(\underline{R})$ requires minimization of $\underline{R}^2$. Since $\underline{R}$ and $\underline{R}^2$ also determine

---

[3]For logical reasons a utility function, over a given range of possible outcomes, can only be approximated by a quadratic function whose maximum is greater than the highest possible outcome. Otherwise we are led to the intuitively unreasonable conclusion that, at some point, utility is an inverse function of the size of the return.

the mean and variance maximization of $\underline{u}(\underline{R})$ requires a preference for low variance. The experimental task permitted choices only of decisions with the same mean. It follows that the preferred choice, in terms of maximization of the utility function approximated by equation 8, should be the investment decision with minimum variance.

The distribution of capital over $\underline{k}$ investments with varying degrees of probability of payment p_i, which will result in minimum variance for the probability distribution of returns can be determined. Two degenerate cases should be noted. If one alternative always pays (no risk) the minimum variance solution is to concentrate all capital on it. If there are two independent sets of alternatives, (1) and (2), such that

$$9. \quad p_i y_i^{(1)} = p_i y_i^{(2)} \quad \text{for all } \underline{i}$$

the minimum variance solution is to allocate equal resources to equal risk investments in each set. These two cases appeared in the experimental situation.

Method

Subjects were asked to recommend a portfolio for a company which wished to obtain income from surplus capital funds. The subject was free to recommend any distribution of a fixed sum ($30,000) over any of sixteen specified bond issues. All money had to be invested. Each bond was described in terms of cost per bond, interest rate, and probability that payments could be met over the ten year life of the bond. Four levels of probability of payment were used; 1.0, .9, .6, and .3. These will be referred to as risk levels (risk = one minus probability of payment). Interest rates were adjusted for the risk levels so that the expected annual profit per dollar invested was .054 for all bonds. Four bond prices ($20, $40, $60, and $80) were used at each risk level. Fictitious names were assigned to the issues. These were chosen so that the nature of the issuing company could not be inferred from its name (e.g. "German and Spezio"). No company bearing any of the names used is known to the author.

Subjects. Thirty-two advanced undergraduate economics students in a large private Eastern university served as subjects. Both men and women participated.

Procedure. The experimenter was introduced to the subjects by their instructor during the last half of their regularly scheduled class period. The experimenter and an assistant then handed out booklets containing the instructions and the experimental material. Some (randomly selected) students left the room

to participate in a related experiment (Hunt and Rowe, 1960). The following instructions were given:

> As an investment counselor for a corporation, you have been asked to invest $30,000 of the corporation's surplus funds in bonds. The Board has indicated that it is interested in the bonds listed on page 2. Your Research Department has given you the market price for each bond, the income of the bond last year, and the department's best estimate that the bond will give this annual income over a period of years. This information is listed on page 2.
>
> For the purposes of this study you should assume that:
>
> (1) In the long run, each bond will remain at its present market value;
>
> (2) In any given year, it will either pay the income stated below or pay no income at all;
>
> (3) In the long run, whether or not it will pay is stated accurately in the probability values listed with each bond.
>
> Study this information carefully and then fill out the Order Blank, indicating how you advise the Board to invest its surplus funds. You may invest in as many or as few companies as you wish. In the space provided on the Order Blank state as precisely as possible how you arrived at your decision.
>
> Please do your figuring on the paper provided.

Subjects were allowed approximately fifteen minutes to complete the experimental task.

Results and Discussion

The relation between average amount invested and risk level is presented in Table 1. An analysis of variance demonstrated a significant difference between the amounts invested at each risk level ($p < .001$). The <u>average</u> trend suggests that subjects prefer to diversify and invest in low risk alternatives. However a closer analysis of the data reveals some paradoxes. Since subjects diversify and display systematic preferences the utility curve, as represented by the average investment, is neither linear nor convex. But if subjects were maximizing a concave utility function they would concentrate all their capital on the risk-less alternative. This is the preferred, but not the sole, choice.

Strictly speaking, the rank order determined by mean amount invested at each risk level does not agree with that predicted by a theory of minimization of a variance, which is equivalent to maximization of a concave utility function. This is true even if one makes the intuitively reasonable assumption that a stated probability of 1.00 is interpreted as "almost no risk." There is a higher average investment at the highest risk level (probability of payment of .3) than the next

Table 1

Average Amount of Money Invested in Each Alternative

Probability of Payment	Order of preference at risk level 1	2	3	4	Total
1.0	8,002	3,271	1,599	505	13,377
.9	5,630	3,153	1,250	483	10,516
.6	1,672	598	325	208	2,803
.3	1,802	722	405	375	3,304

highest risk level (probability of payment of .6). The difference is not statistically significant; however it cannot be maintained that the lowest risk alternative is preferred for every possible pair of alternatives.

The utility maximization hypothesis also fails to predict individual patterns of investment. An approximation to utility maximization is to require that a subject must invest no more money in any given investment than he does in investments at a lower risk level. Only 17 of the 32 subjects exhibited this weak rank ordering over risk levels. This figure includes two subjects who stated that they could not discriminate between the investments. Although such an answer can be interpreted as satisfying the weak order approximation of a concave utility function it seems more accurate to say that it satisfied the requirements of a linear utility function.

Subjects might be maximizing expected utility calculated with subjective probability. In this case the expected monetary profit would be a function of the (unknown) relation between subjective and objective probability. It would then be impossible to interpret investment preferences. However it does not appear that a subjective probability function distorted the results of this experiment. Examination of the subjects' "scratch sheets" showed that virtually every subject made explicit use of the stated probability of payment in calculating the expected value of each investment. Only two subjects did not calculate expected values correctly; their mistakes were clearly due to arithmetical error. They were dropped from the experiment.

A further paradox appears when the distribution of capital over investments

at the same risk level is examined (Table 1). Subjects do not concentrate their capital on any one investment within a risk level. On the other hand they do not distribute their capital uniformly over alternatives at the same risk level. Uniform investments do appear more frequently at high risk levels. A concave utility maximizer would be more concerned with the distribution of investment over several uniform alternatives at high risk levels than low since the difference between concentration and diversification strategies, in terms of expected utility, is a direct function of the risk level.

Latané (1960) has reported an experiment in which he obtained similar results. Students in an investment class acted as advisors to a hypothetical individual investor. They were required to make paired discriminations between a standard "no risk" stock and various risk alternatives. Their choices were not completely consistent with preference for high expected returns or low standard deviation of return. In addition Latané found that subjects' choices did not always maximize the geometric mean of the probability distribution of returns. Latané (1959) has shown that this is a rational strategy when profits are to be reinvested and (following an argument first presented by Bernoulli) when maximization of a concave utility function is a goal. Latané's experimental results suggest that risk variations have a psychological contribution beyond their effect on expected utility.

In the light of these experiments on recommended choice, what is the status of the idea of maximization of utility as a descriptive theory? Perhaps a clear cut utility function would have been revealed if subjects had to make decisions concerning their own money and/or real transactions. While it is true that the results of such experiments have yielded individual utility functions (Mosteller and Nogee, 1951; Davidson, Suppes, and Siegel, 1957) such results have their own limitations in terms of amount of money. In spite of the lack of personal "stake" in the decisions, role playing by economics students does have some face validity. Behavior in real betting situations with small amounts of money and role playing experiments should both be investigated. In addition there is a need for well controlled field studies of economic decision making. Such studies may demonstrate advantages and disadvantages of each experimental method.

To the extent that the present results are a valid test, utility maximization as a descriptive theory is open to question. A theory of economic decision making does not have to involve maximization of a continuous function of the pay-off commodity. A theory which separates alternatives into discrete sets, or which

makes a distinction between investment and gambling, or which permits simultaneous operation of more than one decision rule appears to be at least worth exploring.

Summary

Economics students were asked to recommend investments over alternatives with equal expected values and varying degrees of risk. Their choices were contrasted to choices expected from maximization of utility. Concave, convex, and linear utility functions were considered.

When averaged over subjects the results were most closely approximated by a theory of maximization of a utility function which is concave in money. However, individual investment patterns were widely varied and did not conform to a theory of maximization of any of the three utility functions considered. Discrepancy from the utility maximization principle was also observed in the average data. It was suggested that although utility maximization may be useful in predicting average behavior a different type of theory is needed to describe individual economic decisions.

References

Davidson, D., Suppes, P., and Siegel, S. Decision Making, An Experimental Approach. Stanford, Calif.: Stanford U. Press, 1957.

Edwards, W. The theory of decision making. Psychol. Bull., 1954, 51, 380-417.

Friedman, M. and Savage, L.J. The utility analysis of choices involving risk. J. Political Economy, 1958, 56, 279-304.

Hunt, E.B., and Rowe, R.R. Group and individual economic decision making in risk conditions. See pp. 21-25 of the present report.

Latane, H.A. Criteria for choice among risky ventures. J. Political Economy, 1959, 67, 144-155.

Latane, H.A. Individual risk preference in portfolio selection. J. Finance, 1960, 15, 45-52.

Markowitz, H.M. Portfolio Selection. New York: Wiley, 1959. (Cowles Foundation Monogr. No. 10).

Mosteller, F. and Nogee, P. An experimental measurement of utility. J. Political Economy, 1951, 59, 371-404.

Suppes, P. and Walsh, K. A non-linear model for the experimental measurement of utility. Behavioral Science, 1959, 4, 204-211.

GROUP AND INDIVIDUAL ECONOMIC DECISION MAKING IN RISK CONDITIONS

Earl B. Hunt[1]

and

Richard R. Rowe

Almost any plan or policy may be thought of as a strategy for decision making. The decisions are often made under risk conditions, when one of several alternatives must be chosen, there being associated with each alternative a set of probabilistically determined rewards. The pure risk condition exists when the rewards and associated probabilities are known for every alternative. When the rewards and probabilities are not known an element of uncertainty is introduced. The decision making process may be broken down into two steps. First an attempt is made, through gathering and evaluating information, to reduce the element of uncertainty. When the problem has been reduced to an acceptable approximation to a decision under risk, a rule for decision making under risk is applied. While it is possible to apply decision rules developed for decision problems under uncertainty, we maintain that the two step procedure is often followed. In particular many organizations exist largely for the purpose of evaluating information and then recommending a course of action. For example, these are major duties in stock brokerage firms.

The structure of a decision making organization may affect its decision making processes. This could happen either in reducing the problem under uncertainty to one under risk or in solving the problem under risk. The latter case is of particular interest in the development of a descriptive model of decision making. If a person asks for advice on the same problem from two different types of organization he should, presumably, receive the same advice. Given constant resources and motives the solution to a decision problem should be a function of the problem and not of the person computing the solution. But whether this is psychological fact can be questioned.

In the simplest organizational dichotomy, will groups and individuals, given the same information, offer the same solution to a decision problem under risk? To avoid trivial agreement the problem must be one which does not offer

---

[1]During the period in which this work was performed the first author held General Electric or National Science Foundation graduate fellowships.

an easy, clearly correct solution (such as an easily recognizable means of maximizing profits in the reward commodity and any reasonable utility function of it). On the other hand the problem must not be so hard that a spurious "group superiority due to group climate" effect appears. Such an effect could be due either to the longer time the group has to work on the problems in man-hours or due to the fact that the probability that a group contains a single gifted individual is greater than the probability that a single individual is gifted (Taylor and McNemar, 1955).

In order to test for the existence of differences in economic decision making in the simple organizational dichotomy of group vs. individual we examined hypothetical portfolio recommendations made by individuals and ad hoc groups. Information about alternative investments was presented so that it was equally available and easy to evaluate. Thus differences between the two recommendations could be attributed to group and individual decision procedures.

Method

Task. This study was performed in conjunction with a study of individual decision making under risk (Hunt, 1960). Subjects "role played" as investment advisors to a hypothetical company. The company wished to invest $30,000 in surplus funds in one or more of 16 bond issues. These issues varied in interest rate and probability that the issue would pay at all. These two variables were adjusted so that the expected annual profit for any distribution of capital was 5.4¢ per dollar. The bond issues were classified by probability of payment: 1.0, .9, .6, and .3. The task is described in detail in the report of the related study (Hunt, 1960).

Design. The task provided four levels of risk (1 minus probability of payment). Subjects were assigned randomly to the individual work condition or to one of ten groups of three. Thus any recommendation could be examined for the effects of risk and method of decision making.

Subjects. Sixty-two men and women students in two advanced undergraduate economics classes were used as subjects. Subjects were assigned to the individual or group conditions by random selection within each class. Two subjects in the individual condition stated that they could not discriminate between the alternatives presented. These subjects were not included in the group vs. individual analysis.

The choice of economics students as subjects gives some face validity to our results. The students understood the task. They were interested in our inquiry into psychological aspects of economic behavior. Also, it is not unreasonable to suppose that some of our subjects will face real-life tasks requiring them to give investment advice.

Procedure. Twenty minutes before the end of a regularly scheduled class period the instructor introduced the experimenter, who read the instructions to the subjects while an assistant passed out the experimental material. Individual and group instructions were identical except that the phrase "work as a group" was substituted for the phrase "work individually."

The subjects in the "individual" condition remained in the class room. Subjects in the "group" condition were escorted to a large room where they could see but not hear other groups. In each condition the subjects were allowed 15 minutes in which to decide upon their recommendations.

Results

Analysis of variance techniques were applied to determine the differences between recommendations. The mean scores for individuals and groups are presented in Table 1. Both individuals and groups preferred low risk investments ($p < .001$)

Table 1

Average Amount of Money Invested in Each Alternative

Probability of Payment		Order of preference at risk level 1	2	3	4	Total
1.0	group	6,188	3,687	2,088	2,037	14,000
	individual	8,002	3,271	1,599	505	13,377
.9	group	3,062	2,563	1,062	1,062	7,750
	individual	5,630	3,153	1,250	483	10,516
.6	group	1,163	1,012	862	713	3,750
	individual	1,672	598	325	208	2,803
.3	group	1,788	1,537	638	537	4,500
	individual	1,802	722	405	375	3,304

and did not differ in the degree of this preference (F for conditions x risk interaction < 1.00). The groups demonstrated a slight tendency to diversify their investments within a risk level more than the individuals. However this trend is only suggestive (p. for conditions x order of concentration < .10).

Discussion

This experiment revealed little, if any, difference between groups and individuals in their methods of choosing alternatives under risk. The suggestive difference in diversification appears, upon examination of individual protocols, to be due to the behavior of a few individuals who, after having invested most of their capital in low risk bonds, "take a flyer" by placing some money with a single, high risk issue. No group adopted such a strategy. Written comments by the subjects who made the single, high risk investments suggest that they regarded their moves as gambles after having protected the majority of their capital. One might conjecture that the group setting discouraged such behavior; however the conclusion cannot be drawn on the basis of this data. Since the majority of individuals do not follow a combined investment and gambling strategy they would, on the average, be expected to dominate groups and force a pure investment strategy. A larger experiment is needed to demonstrate the reliability of this difference and to determine whether groups dominated by "risk takers" appear more or less frequently than would be expected by chance.

This study suggests that after a decision problem under uncertainty is reduced to one under risk both individuals and groups will, on the average, apply the same decision rule. An analysis of individual choices (Hunt, 1960) showed that this rule approximated maximization of a utility function whose relation to profit was concave upward. At certain points the approximation breaks down.

Summary

Individuals and three man *ad hoc* real groups recommended investments in alternatives differing in risk and amount of payment but identical in expected monetary profit. No reliable differences were found, suggesting that individuals and groups apply the same decision strategies to choices in the risk condition.

References

Hunt, E.B. Maximization of utility in economic decisions under risk. See pp. 12-20 of the present report.

Taylor, D.W. and McNemar, Olga. Problem solving and thinking. In Stone, C.P. (Ed.), Annual Review of Psychology. Palo Alto, Calif.: Annual Reviews, Inc., 1955, 6, 455-482.

INFORMATION SEEKING IN SEQUENTIAL DECISION MAKING AS DEPENDENT UPON TEST ANXIETY AND UPON PRIOR SUCCESS OR FAILURE IN PROBLEM SOLVING

John S. Roberts, Jr.

Little experimental work has been done on sequential decision making. This is true in spite of the fact that there are many situations in which individuals must engage in such decision making. In these situations, the individual at each step in the process may either choose among the alternatives available to him or decide to obtain more information before making a choice. Moreover, the situation may often be such that the longer he waits to make a final choice, the smaller will be the payoff for selecting the best alternative. Hence, he may often be faced with a conflict between deciding early with the probability of finding later that he made the wrong choice and waiting until he has more information and is surer of his choice.

A number of studies have been done on simple decision making. In an experiment by Winder (1953), for example, subjects were asked to make psychophysical judgments. Individual differences in making such judgments were related to certain personality variables, including ego control and appropriateness of interpretation of ink blots. Johnson (1954), in an experiment using measures of speed and judgment in decision making, found no relation between these variables and the personality measures of manifest anxiety, ego control, authoritarianism, and need achievement.

In a study of individual differences in sequential decision making, Pruitt (1957) obtained a measure of the amount of information required by subjects before making a decision in each of four different kinds of problems. In two of the problems the individual had to decide, on the basis of information provided by a series of red and green lights, which of two alternative conditions a machine was in. In the first problem the individual began by choosing one of the alternatives and then had to decide, on the basis of the series of lights, whether to switch to the other alternative. In the second problem he made no initial choice but rather decided which condition the machine was in only after he had seen some part of the series of lights. In both he was told that the longer he took to make a decision, the fewer points he would receive for being correct. The other two problems employed required the subject to decide on the basis of a series of slides which of two lines was the longer. In one the

subject was offered no incentive for deciding early or deciding correctly. In the other, the subject was told, as in the first two problems, that the longer he took, the fewer points he would receive for being correct. The intercorrelations among the four types of problems in amount of information taken were all significant and ranged from .35 to .76. Pruitt's finding which is of most interest in the present context resulted from the use of a questionnaire designed by him to measure level of manifest anxiety. Positive correlations of .54, .57, .36, and .58, respectively, were obtained between scores on this questionnaire and amount of information required for decision in each of the four types of problems described above. The higher the level of anxiety, the great the amount of information sought before making a decision. It is of interest to note that the lowest correlation, .36, was obtained with the only problem in which the instructions were intended to provide no incentive for deciding early or correctly. This suggests that one effect of incentive may be to enhance the relation between anxiety and information seeking.

The purpose of the present study was to explore further the relation between level of anxiety and information seeking in sequential decision making. One part of the study was designed to determine whether the correlation between anxiety and information seeking found by Pruitt would be obtained if a different test of anxiety and a different sequential decision task were employed.

The study was also designed to test a second hypothesis. The correlation reported between individual differences in anxiety and individual differences in information seeking suggests that if anxiety level were changed experimentally, a corresponding change in information seeking would be observed. The assumption was made that anxiety could be increased experimentally by providing subjects with an experience of failure or reduced by providing them with an experience of success. The specific prediction was that subjects who have just experienced failure in attempting to solve a series of problems will seek more information in subsequent sequential decision making than will subjects who have just experienced success in attempting to solve an equal number of similar problems.

Procedure

Several tests of anxiety were considered for possible use, including the Taylor Manifest Anxiety Scale, the Mandler-Sarason Test Anxiety Scale, and others. The one finally selected for use was the Achievement Anxiety Test

constructed by Haber and Alpert. Data obtained by them show that scores on it are well correlated with scores on other specific anxiety tests. The test has the important advantage of being brief. It also has another feature which it was thought originally might be important; it is designed to provide both a measure of debilitating anxiety and a measure of facilitating anxiety.

Two sets of ten problems each were selected from among a much larger number of problems employed in previous experiments. The problems employed included primarily spatial and arithmetic reasoning problems. The two sets of problems appeared quite similar, but they differed in one important respect. One set included problems which data from previous experiments had indicated would be very difficult for college undergraduates to solve within the time of three minutes provided for working on each one. The other set included problems which such data had indicated would relatively easy for the same subjects to solve within the same time limit. In a pretest with 18 students not included in the experiment, no one solved more than four of the ten in the first set and no one solved less than seven of the ten in the second test.

The following two problems are from the difficult set:

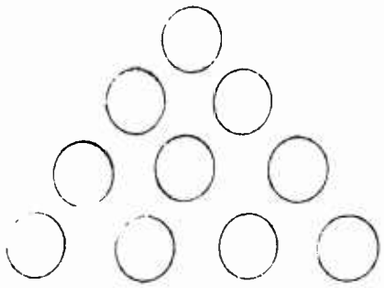
> How can you bring up from the river <u>exactly</u> six quarts of water when you have only a four and a nine quart pail?

> Imagine that you have a piece of cardboard of the following size and shape. Draw lines to show how it could be cut into four smaller pieces all of the same size and shape.

The following two problems are from the relatively easy set:

> Snuff, the tramp, rolls his own cigarettes from butts he collects in his travels. The tobacco from six butts produces one new cigarette. One day he collected a total of 72 butts. He smoked a cigarette every half hour, yet this supply lasted him seven hours. How did he manage this?

The triangle below is made up of 10 pennies. Show how you could move only 3 of the pennies to turn the triangle upside down—make it point down instead of up.



Three sequential decision making tasks were employed. Each involved presenting the subject with a set of clues one at a time and asking him to the object or animal being described. This task is somewhat like the well-known game of Twenty Questions which has been previously employed in experimental studies of problem solving (Taylor and Faust, 1952). In the present task, however, the subject does not ask his own questions but instead receives a series of clues presented in a fixed order at the rate of one every ten seconds.

The tasks were constructed so that insofar as possible information relevant to the identification of the object or animal would be knowledge common to all subjects. This was done, of course, in order to minimize any possible differences among subjects in number of clues taken resulting from differences in relevant knowledge. The tasks were also constructed so that a unusually large number of clues would be required before the subject could with a high degree of probability identify the object or animal being described. The purpose was to make number of clues taken as sensitive as possible as a measure of differences in information seeking. Tables 1, 2, and 3 present the three Clues Tasks employed.

The Achievement Anxiety Test was administered by the experimenter to 65 students at Yale in two classes in introductory psychology. It was announced at the time of the administration of the Test in each of the two classes that there would be a second session of the study in which members of the class might be asked to participate. Instructions for the Test emphasized that all answers would be held in confidence and that the questionnaire in no sense involved competition among those filling it out.

Table 1

First Clues Task

1. Animal
2. Living
3. Vertebrate
4. Non-human
5. Its flesh is edible
6. Does not fly
7. Does not swim
8. Mammal
9. Four-legged
10. Covered with hair
11. Larger than a cat
12. Has hooves
13. Makes characteristic vocal noises
14. Eats grass
15. More often found in European countries than in this country
16. Can be domesticated
17. Often found on farms
18. Its skin is useful for making products
19. Not a cow
20. Smaller than a horse
21. Has a short tail
22. One species found in mountainous regions
23. Has horns
24. Not a sheep
25. The milk of the female can be consumed by humans
26. The female's milk is used to make a particular kind of cheese
27. Inclined to butt objects with its horns

ANSWER: GOAT

Subjects were contacted about five weeks after they had filled out the anxiety questionnaires in class and asked to participate in an individual experimental session. They were told that they would be paid $1.25 for about 50 minutes of their time. They were given no information about the experiment except that it was related to the first questionnaire and that it would be of the pencil and paper variety. Of the 65 students who took part in the first session, 44 participated in the second session. The reason for the reduction in number was that some of the 65 had in the interim taken part in another experiment employing similar tasks (Worley, 1960) and some were simply not interested.

In the second session, each subject was first given one of the two sets of ten problems to solve. A random procedure was employed in determing

Table 2

Second Clues Task

1. Vegetable
2. Non-living
3. An object
4. Manufactured
5. Able to be lifted and carried
6. Always made in the same basic shape
7. No moving parts
8. Can be used by both men and women
9. Used for recreational purposes
10. Held while being used
11. Used in a particular sport
12. Used seasonally
13. Comes in contact with another object when used
14. Is not thrown
15. Does not roll
16. Not any sort of ball
17. Sport in which it is used is played outdoors
18. Longer than it is wide
19. Made of wood
20. Made in standard sizes
21. Not any sort of racquet
22. Made from a single piece of material
23. Not a hockey stick
24. Round in one dimension
25. Tapered
26. Made by being turned on a lathe
27. Sport in which it is used is played in spring and summer
28. Sport in which it is used recently added a third league

ANSWER: BASEBALL BAT

which 22 of the 44 subjects would receive the difficult set and which 22 the easy set. All subjects were instructed in part:

> You may remember that we had you fill out a questionnaire concerning your attitudes toward exam taking. This time we want to see how well you do under real test conditions.
>
> In this booklet is a set of ten problems of the berbal and spatial kind. They are a type which should be familiar to you. None of these are trick questions--they all have answers.
>
> We are using this particular test because it includes problems which have been given to large numbers of college students like yourself and will give us a good indication of how you compare with others of your general intelligence level. We have found that the average college student can solve six or seven of the ten problems in the allotted time.

Since it was expected that in fact almost none of those given the difficult set,

Table 3

Third Clues Task

1. Vegetable
2. Non-living
3. An object
4. Manufactured
5. Able to be lifted and carried
6. Used by both sexes
7. Used by all age groups
8. Mainly a wood product
9. A paper product
10. Printed matter
11. A number of these are printed at the same time
12. Made with a variety of designs
13. Produced in nearly all countries of the world
14. Often made with more than one color
15. Not any sort of book
16. Weighs less than a pound
17. Not any sort of newspaper or magazine
18. Essential to the performance of a service
19. Not any sort of ticket
20. Has more value than cost of materials
21. Comes in different denominations
22. May be quite valuable
23. In United States is made under the supervision of the government
24. Many types commemorate some person
25. Not any form of money
26. Comes in various sizes
27. Many types commemorate some event
28. Usually every dimension less than two inches
29. Has glue on unprinted side
30. Its edges are perforated

ANSWER: POSTAGE STAMP

and almost all of those given the easy set, would solve "six or seven of the ten problems," the intent of these instructions was to make one group experience failure and the other group experience success.

The subjects were also told that they would have three minutes for each problem and that upon signal they must go on to the next problem; they were given the option of going on to the next problem if they obtained a solution that they were certain was correct in less than three minutes; at no time could they return to an earlier problem. To increase motivation, they were told that they could find out later how well they had done.

The three sequential decision tasks were administered to each subject immediately following the completion of the set of ten problems. The nature of the Clues Tasks was explained to the subject and he was instructed in part:

> Try to make the best score you can. Your score will depend on the number of clues it took you to get an answer; that is to say, the fewer clues you use the higher your score will be. However, be reasonably sure your answer is correct as you may give only one answer and you will receive no credit for wrong answers. In other words, your score will depend both on accuracy and on using as few clues as possible.

Each clue was typed on a 4" x 6" card. The subject was instructed to turn to the first card and read the clue when told by the experimenter to begin. At the end of ten seconds, the experimenter said "Next" and the subject turned to the next clue. He continued to turn cards at ten second intervals until he made a decision as to what the object being described was.

When the subject had completed all three tasks, he was told the real purpose of the experiment, paid, and asked not to discuss the experiment with others. If he wished, he was shown the correct answers to all problems.

Results

The Achievement Anxiety Test yielded two scores for each subject, one for debilitating and the other for facilitating anxiety. However, since for the 65 subjects a correlation of -.55 was found between these two scores, it was decided to obtain for each subject a single composite debilitating score by combining appropriately the original two scores. The resulting composite scores had a mean of 40.3 and ranged from a low of 30 to a high of 70.

An analysis of the performance on each of the two sets of ten problems showed that the intended results were obtained. The mean number of problems solved by the 22 subjects given the difficult set was 2.18; only 3 of the 22 solved as many as four problems and none solved more. The mean number of problems solved by the 22 given the easy set was 7.91; only 3 of the 22 solved less than seven problems and none solved less than four. Clearly, one group may be said to have experienced failure and the other success in relation to what they had been led to expect they should achieve.

Table 4 presents the correlations obtained between the composite scores on the Achievement Anxiety Test and the number of clues required for decision in each of the three task. None of the correlations differed significantly from

Table 4

Correlations between Scores on Achievement Anxiety Test and Number of Clues Required for Decision

	Task		
	Goat	Rat	Stamp
Success	-.27	-.08	-.34
Failure	.25	-.30	.17

zero. Hence, the first prediction is not confirmed.

The data in Table 4, however, suggest an interesting and unexpected possibility, namely, that the relation between individual differences in anxiety and information seeking under success may differ in direction from that under failure. Of the three correlations under success all are negative, whereas two of the three under failure are positive. When for each task the difference between the coefficients under success and failure was examined by employing the z transformation and a t test, two approached significance, p being about .10 for the first and the third task.

In the next analysis, both the 22 subjects who had experienced success and the 22 who had experienced failure were divided into two groups. On the basis of their scores on the Achievement Anxiety Test, 10 of the 22 were assigned to the "high anxiety" group and the remaining 12 to the "low anxiety" group. The reason for not dividing the 22 into two groups of equal size was that in each case this would have resulted in assigning to different groups two subjects who had identical scores on the Achievement Anxiety Test.

Table 5 presents the mean number of clues taken for each task under conditions of success and failure for both high and low test anxiety. In five out of the six cases the results were in the predicted direction, i. e., subjects who had experienced success used less information than did those who had experienced failure.

To test the significance of this finding, an analysis of variance was carried out (see Table 6). The results of this analysis showed that the effect of the success-failure variable was significant at the .05 level, thus confirming the second prediction. The only other F-ratio which approached significance was that for the difference among tasks, a difference of no theoretical interest.

Table 5

Mean Number of Clues Taken for each Task under Conditions of Success and Failure for High and Low Test Anxiety

		Task		
		Goat	Bat	Stamp
High Anxiety	Success	16.6	16.0	17.1
	Failure	18.7	14.8	21.6
Low Anxiety	Success	17.1	16.0	18.8
	Failure	17.8	20.3	20.1

Table 6

Analysis of Variance

	d. f.	Mean Square	F-Ratio	p
Between Subjects				
Test Anxiety	1	36.04		
Success-Failure	1	126.95	3.50	.05*
Test Anxiety x Success-Failure	1	6.20		
Error	40	36.27		
Within Subjects				
Tasks	2	73.73	2.85	.10
Tasks x Test Anxiety	2	9.39		
Tasks x Success-Failure	2	5.89		
Tasks x Test Anxiety x Success-Failure	2	25.35		
Error	80	25.84		

*A one-tailed test was employed since the direction of the difference was predicted.

Discussion

The present study was designed in part to determine whether the positive correlation between anxiety and information seeking would be obtained when a different measure of anxiety and a different sequential decision task were employed. The data obtained failed to confirm Pruitt's finding. None of the six correlations between scores on the Achievement Anxiety Test and number of clues taken in any of the three Clues Tasks differed significantly from zero. Question may be raised concerning the test of anxiety or the decision task employed here. However, as previously noted, data are available (Haber and Alpert) which show that scores on the Achievement Anxiety Test correlate well with scores on other widely used measures of anxiety. Moreover, data are also available (Worley, 1960) showing that information seeking on the type of decision task employed here correlates significantly with information seeking on two other types of sequential decision tasks. The explanation of the difference between the present finding and that of Pruitt (1957) must await further work.

Although not significantly different from zero, all three of the correlations between anxiety and information seeking under success were negative, whereas two of the three under failure were positive. The difference between the coefficient under success and that under failure approached significance for two of the tasks. These unexpected results can be regarded only a suggestive. Nevertheless, they suggest strongly that in future studies the possibility should be explored that the relation between individual differences in anxiety and in information seeking under success may differ from that under failure. The explanation of the negative correlations under success is puzzling. If true correlations, they indicate that the higher the level of anxiety, the less the information sought; why experience of success should foster such a relation is not immediately apparent.

The present study was also designed to test the prediction that subjects who have just experienced failure in problem solving will seek more information in subsequent decision making than will subjects who have just experienced success. This prediction was based on the assumption that failure will tend to increase anxiety and success to reduce it. The data obtained confirm the prediction. Since the differences obtained were significant only at the .05

level, replication of this finding would be desirable. It is paradoxical that this prediction that experimentally-produced failure would increase, and success would reduce, information seeking was based on the expectation of a positive correlation between individual differences in anxiety and information seeking. The prediction was confirmed, but the expectation upon which it was based was not. Explanation of the paradox perhaps should await confirmation of present findings correlation of individual differences.

References

Alpert, R. and Haber, R.N. Anxiety in academic achievement situations. Unpublished memorandum.

Johnson, L.C. Speed and Confidence of Judgment as Psychological Variables. Stanford: Dept. Psychol., Stanford Univer. (Tech. Rep. , Contract Nonr 225-01).

Pruitt, D. An Exploratory Study of Individual Differences in Sequential Decision Making. Ph.D. Dissertation, Yale Univers., 1957.

Taylor, D.W. and Faust, W.L. Twenty questions: efficiency in problem solving as a function of size of group. J. exp. Psychol., 1952, 44, 360-368.

Winder, C.L. Decision Making. Stanford: Dept. Psychol., Stanford Univer., 1953, (Tech. Rep. 1, Contract Nonr 225-01).

Worley, D.R. Amount and generality of information-seeking behavior in sequential decision making as dependent on level of incentive. 1960. See pp. 1-11 of this report.

TWO EXPLORATORY STUDIES OF THE EFFECT OF SEPARATION OF PRODUCTION FROM EVALUATION OF IDEAS

David L. Singer

The suggestion has repeatedly been made that creative thinking is facilitated by separating the process into first a stage of production of ideas without criticism followed by a stage in which the ideas are evaluated. It was this hypothesis that the present two experiments were designed to explore.

In an informal memorandum on creativity, Miller has advocated this method as being one which reduces fear during thinking. He considers fears, both on a verbal and non-verbal level, to impede the creative process by preventing the individual from thinking freely and producing the wide, free range of verbally mediated responses necessary for originality and creativity. Typical among these fears are: fear of being unconventional, fear of thinking socially unacceptable thoughts, etc. It is his opinion that by deliberately deciding to suspend criticism and judgment such fears may be minimized.

This view seems compatible with the psychoanalytical hypotheses that artists are more able to tolerate thoughts, feelings, and impulses which would arouse anxiety in others. Psychoanalytic theorists have viewed the creative processes as being aided by, and relying upon, a process termed "regression in the service of the ego." The hypothesis formulated by Kris (1952) states that creativity is related to preconscious and unconscious needs and impulses, and their gratification in fantasy. The regression of which they speak is a shift to a more primitive mode of thought called "primary process." Primary process thinking is closely related to the type of thinking which is found in dreaming, and to the less reality-oriented thinking of the very young child. Conscious ideas and percepts are amplified and transformed by unconscious needs and wishes. Such thinking is in the service of the ego when the individual is able to control his regression into the fantasy world and come back to reality-oriented thinking at will. Kris refers to this process as the "inspirational" phase. The thinker must also remold or shape what he has "brought back with him" into a communicable and refined form. This last stage is referred to as "secondary process" thinking. It follows from this that if two people are equal in intelligence and background, the one who is more able to "regress in the service of his ego"

because of his personality structure will be the more creative of the two. Similarly, if "regression in the service of the ego" can be experimentally facilitated, increased creativity should result. The separation of thinking into a production and then an evaluation phase thus would increase creativity by giving the thinker a structure in which possibilities for "regression in the service of the ego" are maximized.

In a more practical vein, Osborn (1957) originated the now famous technique of "brainstorming" on the assumption that such a separation is helpful. In a brainstorming session, groups of people sit together and as quickly as possible throw out as many ideas and suggestions as they can. During this production stage, criticism is strictly taboo. Taylor et al. (1957) have shown that contrary to Osborn's emphasis on the value of group participation, several individuals brainstorming alone will produce more and better ideas than the same number of people brainstorming together as a group. However, the value of the separation of evaluation from production of ideas needs further exploration.

Two related experiments were designed to test this hypothesis. They differed in the nature of the tasks which the subjects were asked to perform. For reasons to be described below, it was deemed necessary to replicate one of the studies.

In practical use of the method to be tested here, the individual either is instructed or instructs himself to postpone for a time criticism or evaluation of the ideas which he is producing. This procedure has the important disadvantage from an experimental point of view in that it leaves uncertain the extent to which subjects are successful in following such instruction. Even though attempting to do so, some subjects may fail to separate production from evaluation.

For this reason, an attempt was made to devise a procedure which would ensure the separation of production from evaluation. What was involved essentially was the presentation to the subject initially of a task calling for the production of ideas useful in the solution of a problem which was not presented until later. Since during the initial period the subject did not know what the final problem was to be, he could not evaluate ideas being produced in terms of criteria relevant to the final problem.

The problem employed in the first experiment involved the making of sentences using only six specified letters. During the first part of work under the separation condition, the subject was asked only to make as many words as he

could using only the letters provided. Later he was given the problem of constructing as many sentences as he could from the specified letters, using the words already completed as an aid.

The problem employed in the second experiment involved creation of a poem using only 17 specified words. In this task under the separation condition, the subject was asked initially to construct as many phrases or sentences as he could, using only the words provided: nothing was said initially about writing a poem. Only later was he given the problem of creating a poem from the 17 words, using the phrases and sentences previously constructed to help him.

It is fully recognized that the procedures employed here did not involve complete separation of production from evaluation of ideas. In each experiment, the task initially presented under the separation condition provided a criterion for evaluating responses. However, since this criterion was much less restrictive than those implicit in the final problem, it seems certain that the amount of evaluation occurring during the first part of work under the separation condition was much less than that occurring during the first part with subjects working on the final problem from the beginning. To the extent that minimizing evaluation during the initial part of work on a problem facilitates creative thinking, it would be expected that performance would be enhanced under the separation condition.

Experiment Ia

Subjects. The subjects were 76 male Brooklyn College students who were divided randomly a "separation group" including 37 subjects and a "unitary group" including 39. Unfortunately, because they failed to produce usable records, four subjects had to be dropped from the former and two from the latter, leaving 33 and 37, respectively. The randomization was achieved by shuffling together two sets of admission cards and by giving each subject when recruited the card at the top of the stack. Each subject was paid a total of $2.50 for participating in both this experiment and Experiment II.

Both groups were run at the same time in the same building, but in different rooms and by different experimenters. Both experimenters were male and approximately the same age.

Procedure. Each subject was presented with an envelope containing instructions and materials. The "unitary" subjects were presented with the six letters,

"T", "A", "D", "M", "E", "N", and were instructed to make as many different sentences as they could from these letters, using only those letters. Forty minutes was allotted for this.

The Separation Group was presented with the same six letters and asked first to make as many words as they could from the letters without being told anything about sentences. They were given ten minutes in which to do this. Pretesting had indicated that this was the optimal length of time for this part of the task, tending neither to waste the subject's time, nor prevent him from making almost as many words as he could from the letters. At the end of these ten minutes, the subjects were instructed to make as many sentences as they could out of only those six letters, and to use the words which they had already made from them as an aid. They were explicitly told that they were not restricted to these words. They were given 30 minutes in which to do this, making a total of 40 minutes—the same amount of time which the Unitary Group was given.

The Separation Group was interrupted for approximately two minutes while these new instructions were explained, and this most probably broke their set. To equalize for this, the Unitary Group was given a two-minute break after their first ten minutes of work.

In comparing the relative creativity of the two groups, the dependent measure used was the number of sentences produced. This was reasonable since the instructions had stressed quantity and had not mentioned quality at all.

Results. The number of sentences produced in 40 minutes by members of the Unitary Group ranged from 34 to 109 with a mean of 90.8. The number produced in the same time by members of the Separation Group ranged from 29 to 207 with a mean of 103.2. The difference between the means yielded a t of 2.42 and was significant at the .01 level with a one-tailed test, the appropriate test since the direction of the difference was predicted. The data thus appeared to confirm the hypothesis.

Unfortunately, however, question was raised by the fact that six of the subjects in the Separation Group unexpectedly made use of ditto marks in constructing their sentences. To the extent that writing speed may have been important, this may have given them an undue advantage. For this reason, the mean for the Separation Group was recomputed, excluding these six subjects. The new mean of 97.6 was found to be not significantly different from that previously obtained for the Unitary Group. It must be noted, however, that these six subjects might well have been among the most productive in the

Separation Group, even if they had not used ditto marks. Hence, neither their inclusion nor their exclusion appears to yield an entirely appropriate comparison. Accordingly, it seemed advisable to replicate Experiment Ia. Experiment Ib was conducted approximately one year after Experiment Ia.

Experiment Ib

Subjects. The subjects were Air Force personnel attached to the Intelligence Corps and studying at the Yale Institute of Far Eastern Languages. They were recruited by putting up notices offering work as subjects in a psychological experiment for pay of $1.25. Each subject volunteered for one of three days on which the experiment was to be run.

The subjects were run in groups which varied in size from eight to 13. Two groups, one Separation and one Unitary group, were run on each of three days. Although on the first of these days the two groups were run by different experimenters in different rooms, on the second and third days an attempt was made to adequately control possible experimenter and room effects through a counterbalancing. This was achieved by having each of the two conditions run in one room supervised by one proctor on one day, and in a different room supervised by the other proctor on the other day. On both days, for both conditions, instructions were read and questions answered by the same experimenter. The proctors merely maintained order and at several points asked the subjects to draw a line under the last sentence they had produced.

By this counterbalancing, any room or proctor effects should appear with equal strength in both conditions. Any interaction between room or proctor and condition should appear in the interaction term of the analysis of variance.

Procedure. Almost exactly the same instructions and material were used in this experiment as in Experiment Ia. There were only two differences. The first and obvious one was that the subjects were explicitly instructed not to use ditto marks. The second was that for both the Unitary and Separation conditions, five minutes was added to the total time. For the separation subjects, this time was added to the second, or sentence making, phase. The rationale behind this was that if, as the data from Experiment Ia suggested, the separation method is superior, the extra time should serve to increase the difference between the groups.

On each of the first two days, all subjects were requested to avoid any discussion of the experiment with their friends.

Results. Table 1 presents the mean number of sentences per subject in each of the two conditions on each of the three days. Table 2 presents the analysis of variance. It will be noted that the absolute difference between conditions is almost exactly the same for each of the three days; hence, there is no need to present a separate analysis for the counterbalanced 2 x 2 design fo the second and third days.

Table 1

Mean Number of Sentences per Subject

	First Day	Second Day	Third Day
Separation	56.7	87.9	127.2
Unitary	75.5	106.5	147.5

Table 2

Analysis of Variance*

Source	d. f.	Mean Square	F	p
Conditions	1	4,577.7	2.51	.15
Days	2	22,034.2	12.06	.001
Conditions x Days	2	383.6	.21	
Error	51	1,826.5		

*N for the separation condition was 7, 11, and 10 for the three days, respectively with a total of 28. N for the unitary condition was 8, 11, and 10, respectively with a total of 29.

The first surprising fact about the data obtained is that, in contrast to the results in Experiment Ia, the mean number of sentences created by the members of the Separation Group was smaller, not larger, than the mean number created by those in the Unitary Group. However, although this was true on each of the three days, the difference failed to reach significance at the .05 level (Table 2).

The second unexpected finding was that there was a highly significant

difference between days in mean number of sentences produced (Table 2), the number increasing from the first to the second and from the second to the third day (Table 1).

Discussion. Taken together, the results of Experiment Ia and Ib fail to support the hypothesis that separation of evaluation from production of ideas facilitates creativity. This statement, of course, is based on the assumption that the experimental manipulation achieves that separation, an assumption which appears tenable. In Experiment Ia, the data did yield a significant difference favoring separation. The meaning of this difference, however, was rendered ambiguous by the fact that six of the subjects in the separation condition had unexpectedly employed ditto marks and that exclusion of these subjects from the comparison reduced the difference to insignificance. In Experiment Ib, the difference in means, though not significant, actually favored the Unitary Group.

The highly significant increase in number of sentences produced from day to day is puzzling. Since different subjects were employed on different days, no such increase was anticipated. Indeed, the replication of the experiment on three different days was undertaken simply to permit the use of a larger number of subjects than could be obtained on a single day. Two explanations of the increase appear possible. The first is, that since the subjects themselves chose the day on which they were to participate, some unknown factor led the least able subjects to come the first day and increasingly able subjects to come the second and third days. The other explanation is that subjects did not comply with our request not to speak to others about the experiment. This appears doubtful, however; it seems improbable that simply having some knowledge of the general nature of the task in advance would lead to a large increase in the number of sentences an individual would produce. More important, the difference in means between the two conditions was almost precisely the same on each of the three days, being 18.8, 18.6, and 20.3, respectively. The increase in means over days, though puzzling, does not appear to necessitate any modification of the conclusion concerning the comparison between conditions; the interaction between days and conditions was negligible.

Experiment II

Subjects. Experiment II was conducted in the same rooms, on the same evening, and with the same subjects as Experiment Ia. Moreover, in Experiment II

the same subjects were assigned to the Separation Group and to the Unitary Group, respectively, as in Experiment Ia. This duplication of assignment raises some question as to whether one is fully justified in regarding the members of these two groups in Experiment II as random samples from the same population. The members of these two groups differ in the experience which they received during Experiment Ia, a difference which though it seems improbable might have had some impact on their performance in Experiment II.

The reassignment at random of the subjects from Experiment Ia to the two groups in Experiment did not appear to be feasible. The danger existed that subjects who were in the Separation Group in Experiment Ia and were reassigned to the Unitary Group in Experiment II might try to second-guess the experimenter. They might reason that they had been taught a method of working in the first experiment, and that they were now being tested, with a new task, to see if they would use the new method. The danger of this and similar possibilities appeared to more than offset the disadvantage inherent in assigning subjects to the same conditions in both experiments. An incidental advantage of duplication of assignment was that this made possible the correlation of the performance of subjects in the two experiments.

Procedure. The task in this experiment required the subjects to create a poem out of a list of 17 words which was presented to them. Both conditions used the same set of words. To find a list of words which were sure to lend themselves to poetic efforts, and also to provide some sort of criterion against which to judge the final productions, the word list was obtained by taking a short poem by a capable poet and randomizing the word order. The actual poem used was The Bee by Emily Dickinson:

> The pedigree of honey
> Does not concern the bee
> A clover, any time, to him
> Is aristocracy.

Subsequent questionning ascertained that none of the subjects was familiar with the poem.

Having been given this word list, the subjects of the Unitary Group were instructed that each was to make as good a poem as he could from the words, using each word only once, and using only these words. Forty minutes was allotted for the task. At the end of this time, each subject wrote what he considered to be his best effort in the space provided.

Subjects in the Separation Group, on the other hand, were at first instructed

to make as many phrases, sentences, thoughts, or clauses as they could out of these words. They were encouraged to write down whatever grouping of words came to mind, regardless of how preposterous or silly it sounded. To ensure a minimum of inhibition, the subjects were reminded that they had not put their names on any of the experimental materials, and that they would remain anonymous. Fifteen minutes was allotted for this part of the task.

At the end of these 15 minutes, they were instructed to make as good a poem as they could from the 17 words, using the phrases, sentences, thoughts, etc., which they had just made, to help them. While these would give them ideas, they were further told that they were certainly not limited to them, and could make any new word combinations they wished. As was the Unitary Group, they were instructed to use each word once, and to use only these words. Twenty-five minutes was given them for this part, making a total of 40 minutes, the same amount of time the Unitary Group had. As in Experiment I, since the separation subjects had to be interrupted so as to be given their new instructions at the end of the first 15 minutes, the unitary subjects were also given a short break after 15 minutes. When the 40 minutes was up, the subjects wrote their best poem in the space provided.

Scoring. For each subject, three scores were obtained. The first was a measure of quantity, the second of quality, and the third a composite score obtained by multiplying the first two together.

The quantity score was the proportion of the seventeen available words used correctly with a penalty for any words used incorrectly. If a word was used twice or if a word not in the list was used, the number of such words used incorrectly was subtracted from the number used incorrectly before dividing by 17. Thus for example if 11 words were used correctly but one were used twice and also one not in the list were employed, the individual's score would be (11-2)/17 which equals .53. The quantity score, being a proportion, could range from 0 to 1.00.

To provide a measure of quality, the poem produced by each subject was rated on each of three dimensions: (a) aesthetic quality, (b) form, and (c) content. For each dimension, a five-step scale was carefully constructed with values ranging from 0 to 4. The dimension itself and each step on the scale was defined by a series of phrases in order to make the scale as reliable as possible. The three ratings for a poem by a given subject were summed to provide a total quality score which could range from 0 to 12.

Two raters working independently rated each poem without knowledge of to which group any of the poems belonged. All poems were rated on one dimension at a time in order to make the ratings of the three dimensions as independent as possible.

A correlation was computed between the total quality score assigned to each poem by one rater and that assigned by the other rater. The resulting coefficient provides an estimate of interrate reliability. The correlations were .91 for the Unitary Group, .78 for the Separation Group, and .86 over all subjects. The reliability of the ratings was quite adequate.

A composite score was wanted which would reflect both the quality of the poem and the degree to which the subject complied with the instructions to use only the words on the list and to use each word once and only once. To obtain such a composite score, the quantity score was simply multiplied by the quality score. A multiplicative rather than an additive combination of these two scores seemed desirable. With a multiplicative combination, a subject would get a score of zero for a creation which in the opinion of the judges was but a different random order of the words, but which used all of them, each once; with a quality score of zero and a quantity score of one, the composite score would, of course, be zero. Certainly, the score of zero seems more appropriate for such a product than that which would be obtained by an additive combination of the quantity and quality scores. Similarly and appropriately, with a multiplicative combination a brilliant jutaposition of only two words would receive a lower score than with an additive combination which would sum a high score for quality with a low score for quantity.

The composite score employed could range from 0 to 12.

Results. The total number of subjects in this experiment was 76, with 37 in the Separation Group and 39 in the Unitary Group. Included in the former were four subjects and in the latter were two subjects who participated in Experiment Ia but who failed to produce usable records in that experiment.

The mean quantity score for the Separation Group was .905 and for the Unitary Group .863. The difference yielded a t of 1.41 which fails to reach significance at the .05 level.

Table 3 presents the mean quality scores for each of the three dimensions and for the total scores. None of the differences between the mean for the Separation Group and the mean for the Unitary Group approaches significance.

Table 3

Mean Quality Scores

	Dimension			
	Aesthetic	Form	Content	Total
Separation	1.88	1.64	2.09	5.61
Unitary	2.08	1.69	2.09	5.86

The mean composite score for the Separation Group was 5.09 and for the Unitary Group 5.06. Again the difference was not significant.

Correlations were computed between the scores of subjects upon the task employed in Experiment Ia and their composite scores for the present task. The coefficients were .14 for the Separation Group, -.06 for the Unitary Group, and -.01 for all subjects. It appears that the two tasks involve different abilities, at least insofar as performance on them is represented in the scores employed.

Discussion. Clearly, the results of Experiment II, like those of Experiment Ia and Ib taken together, fail to support the hypothesis that separation of production from evaluation of ideas facilitates creativity. No evidence was obtained for any consistent difference between the two experimental conditions employed here. The interpretation of these findings is, of course, contingent upon the acceptability of the assumption that the experimental procedures employed did in fact result in important reduction in evaluation during the first part of work in the Separation Groups. The assumption still appears tenable and the procedures still appear to provide an experimental control not otherwise available, but the failure to find differences between the experimental conditions suggests that it would be desirable to employ either other tasks or designs in further exploration of the primary hypothesis.

References

Kris, E. Psychoanalytic Explorations in Art. New York: International Universities Press, 1952.

Miller, N. E. Notes on Sources of Difficulty in Creative Thinking. Unpublished.

Osborn, A. F. Applied Imagination. New York: Charles Scribner's Sons, 1957.

Taylor, D. W., Berry, P. C., and Block, C. H. Does Group Participation when Using Brainstorming Facilitate or Inhibit Creative Thinking? New Haven: Depts. of Industr. Admin. and Psychol., Yale Univer., 1957 (Tech. Rep. 1 Contract Nonr 609-20)

A NOTE ON THE RELIABILITY OF FIVE RATING SCALES

Donald W. Taylor

In an experimental study of the effect of group participation upon creative thinking when using brainstorming, Taylor, Berry and Block (1957) constructed and employed five rating scales. In that experiment, 12 groups of four men each and 48 individuals were given the same three problems to work on in the same order. All experimental sessions were recorded using appropriate sound equipment, and essentially complete typewritten transcripts were made of the responses of each group and each individual to each of the three problems.

The data were analyzed initially in terms of number of responses produced and in terms of number of unique responses produced. However, detailed examination of the 483 different suggestions made for solution of the Tourist Problem and also of the 513 different suggestions for solution of the Teachers Problem indicated that these suggestions differed in quality with respect to at least three dimensions: feasibility, effectiveness, and generality. Accordingly, five-step rating scales were constructed for use in measuring these three. Inspection of these scales, shown in Figures 1, 2, and 3, provides the best available definition of each of the three variables. The intention was to

Fig. 1. Feasibility Scale

0 Clearly impossible. No known method of attainment. Contradicts known facts or scientific laws.

1 Very doubtful feasibility. Means of attainment quite unclear. Necessary acceptance highly improbable.

2 Feasible but would require very large expenditures of funds, major political or social changes, or major technological developments.

3 Could be carried out with sizable additional funds, with some limited social or political changes, or with minor technological developments.

4 Could be carried out in the near future and with very reasonable effort or expenditure of funds.

construct scales such that the successive steps on each scale would be subjectively equal, each step would be relatively unambigious, and all five steps would actually be used in rating.

Fig. 2. Effectiveness Scale

0 No conceivable contribution to solution of problem. Suggestion impossible of attainment.

1 Very little, if any, contribution to solution of problem.

2 Probably some contribution to solution of problem.

3 Definite minor contribution to solution of problem.

4 Clearly a major contribution to solution of problem.

Fig. 3. Generality Scale

0 So general as to be meaningless; of indeterminate meaning.

1 Highly specific suggestion (or consequence) within a narrow area.

2 Moderately specific suggestions within a narrow area; highly specific suggestions within a broad area.

3 Broad suggestions within a narrow area; moderately specific suggestions within a broad area.

4 Broad suggestions within a broad area.

The 791 different responses made to the Thumbs Problem differed from those made to the other two problems in that they represented anticipated consequences instead of suggested steps for solution. For this reason, only one of the three rating scales constructed for rating responses to the other two problems, namely generality, appeared equally applicable in the case of the Thumbs Problem. However, analogous to feasibility and effectiveness on the other problems were the dimensions of probability and significance, respectively, for the Thumbs Problem. Accordingly, the two scales shown in Figures 4 and 5 were constructed.

All three authors of the earlier report (Taylor, Berry, and Block, 1957) participated in the rating of the responses to the three problems. The first author rated the responses to the Tourists, Thumbs, and Teachers Problems on effectiveness, probability, and generality, respectively; the second author on

Fig. 4. Probability Scale

0 Very highly improbable or clearly impossible.

1 Conceivable, but improbable.

2 Possible.

3 Probable.

4 Highly probable.

Fig. 5. Significance Scale

0 Irrelevant or impossible consequences.

1 Clearly trivial; of no importance.

2 Probably some effect of very limited importance.

3 Minor impact on daily lives; changes in what many people do frequently or in their ways of doing things.

4 Major impact on daily lives of many people.

generality, significance, and feasibility, respectively; and the third author on feasibility, generality, and effectiveness, respectively. Thus, the responses to each problem were rated on three different scales by three different raters, hence presumably increasing the independence of the ratings of the three characteristics. The intercorrelations between the various pairs of ratings for each problem were in fact low, ranging from -.01 to .38.

The possibility of having all three raters rate the responses to each problem on all three scales was considered at the time of the original study and rejected for several reasons. First, a rough check had indicated that the reliability of ratings by a single rater would be adequate. Second, well over 100 additional man-hours would have been required. Third, the major conclusion of the study was well supported by analyses already completed involving number of

ideas produced and number of unique ideas produced; it seemed quite unlikely that supplementary analyses involving the rating scales would yield any important modification in the major conclusion; the difference between the two experimental conditions in mean number of ideas produced was so large that it appeared improbable that any possible difference between the two conditions in average quality of ideas would be sufficient to offset the difference in number; hence, a large additional investment of time and money in obtaining ratings by all three raters on all three scales for all three problems did not appear justified. However, subsequent to the publication of the original study, question has been raised concerning the reliability of the rating scales employed (Cohen, Whitmyre, and Funk, 1959). For that reason, the study reported here was carried out.

Procedure. From the original large group of responses to each problem, an unbiassed sample of about 100 items was drawn by taking from the original master list (Taylor, Berry, and Block, 1957, p. 11) every fifth response to the Tourists and Teachers Problems and every eighth response to the Thumbs Problem. In no case did the mean or standard deviation of the original ratings for the sample of responses differ significantly from that of the original ratings of the population from which it was drawn.

The original ratings were made in September of 1957. In late August and early September of 1959, each of the three original raters employed again for each of the three problems the same scale which he had used originally, this time rating only the sample of about 100 responses in each case. Because each rater had originally rated a total of 1,787 items for the three problems and because just about two years had elapsed since the original ratings were made, there appears very good reason to believe that the correlation between the original ratings and those made two years later provides one acceptable measure of reliability—one not spuriously inflated by any possible memory of the original ratings.

Two of the raters also rated the sample of responses to each problem on each of the two scales which they had not employed in the original ratings. Hence, for these two raters, ratings made in 1959 were available for each of the three scales for each of the three problems. The correlations between these ratings on the various pairs of scales provide the usual measure of interrater reliability.

In making these latter ratings, each rater rated at one time the responses to a single problem on a single scale; these ratings were thus concealed while

he rated the same responses on a second and in turn on a third scale. About one and one-half hours was required for rating 100 responses on a single scale. That the three raters worked quite independently is perhaps symbolized by the fact that, with one minor exception, the ratings in the present study were made with one rater working in New Haven, one in New York, and one in Washington.

Results. The results obtained are shown in Table 1. The first column presents the correlations between the original ratings and those made by the same rater after two years. The second column presents the correlations between ratings made independently by two different raters in 1959.

Table 1

Reliability of Five Rating Scales

	Correlation of Ratings after Two Years with Original Ratings	Correlation of Ratings Made Independently by Two Different Raters
Tourists Problem (N = 97)		
Feasibility	.85	.83
Effectiveness	.82	.63
Generality	.59	.52
Thumbs Problem (N = 99)		
Probability	.74	.64
Significance	.64	.47
Generality	.68	.74
Teachers Problem (N = 103)		
Feasibility	.59	.67
Effectiveness	.41	.59
Generality	.56	.53
Mean (via z transformation)	.68	.64

Discussion. The data presented in Table 1 clearly support the conclusion that the original ratings had fully adequate reliability for the purpose for which they were used. The considerations which underlie this conclusion, however, are somewhat complex and often misunderstood. Hence, it seems appropriate to discuss them in some detail here.

The first point to be emphasized is that the scales employed are, of course, only one of several factors affecting the reliability of the ratings obtained. If others employing these scales fail to obtain reliabilities similar to those reported here, this may be due to: (a) restriction of range of variation in the sample with respect to the variable being rated; (b) insufficient time devoted to making the ratings; (c) inadequate knowledge on the part of the rater of the domain of ideas to be rated; suggestions for a solution to a problem can hardly be rated well by an individual who lacks knowledge of the problem area; (d) inappropriateness of the variables for use with the sample of responses being rated (different scales were needed in the original study for different problems); (e) inexperience or lack of ability on the part of the rater. The fact that the reliabilities reported here were obtained would appear to demonstrate that the scales employed have the potential for yielding reliabilities at least as high. Whether such reliabilities are attained in practice will depend on other factors of the kind just listed.

A second consideration in evaluating the interrater reliabilities reported here is that the coefficients obtained, ranging from .47 to .83 with a mean of .64, are similar to those ordinarily attained with other carefully-constructed scales. A review of the literature suggests that correlations between two raters above .80 are uncommon for single scales employing five or more steps, and that though coefficients between .70 and .79 are more frequent, one must more often work with ratings having reliabilities between .60 and .69, or even lower.

A third and major point is that the reliabilities of ratings must be evaluated in terms of the purpose for which they are to be used. For this reason, general statements that rating scales are reliable or unreliable appear inappropriate. Such statements assume that if interrater coefficients fall above some arbitrary value, e. g, .80, then the ratings are "reliable" and satisfactory for use and if, on the other hand, the coefficients fall below that value the ratings are "unreliable" and not satisfactory for use. However, whether or not ratings with a given interrater coefficient may be satisfactorily used depends not only upon the size of the coefficient but also upon the nature of the use to which they are to be put.

Intelligence tests provide a familiar illustration of the importance of considering the use to which an instrument is to be put in evaluating its reliability. If one wishes to use the scores obtained to discriminate among individuals, then, of course, a test is needed with just as high a reliability as possible, preferably above .90. If, however, one wishes only to determine whether two groups differ in mean intelligence, than a shorter or more easily-administered test of lower reliability may be satisfactory.

In the original study (Taylor, Berry, and Block, 1957), ratings were obtained employing the present scales, not to be used in discriminating among individuals, but only to determine whether there were significant differences between the two experimental conditions in mean scores based on such ratings. This fact must be kept in mind in assessing the magnitude of the coefficients reported in Table 1.

A fourth and important point stems from the fact that the coefficients reported here represent the reliability of the ratings of ratings of single responses. These coefficients are high enough so that the use of the scales for the purpose for which they were employed would be justified even if the quality score for a given real or nominal group involved only a single rating. This, however, was not the case. A given group produced not a single response to a given problem, but a large number—the mean number of responses to each of the three problems by the 12 real groups was, for example, 37.5. The score which a given group received for a given measure of quality on a given problem—e. g., for generality for the Tourists Problem—represented not the rating of a single response to that problem, but rather the sum of the ratings of all of the responses by that group of four subjects to that problem. Hence, the reliability of the appraisal of the performance of that group on that problem would be considerably higher than estimate of the reliability of single ratings reported in Table 1. Just as the total score on a test is more reliable than the score on any single item, so would an appraisal for a single group obtained by summing the ratings for a large number of responses be more reliable than the rating for a single response. Hence, the correlations reported here underestimate the reliabilities of the appraisals of the performance of single groups actually employed in the original analyses.[1]

---

[1] No attention has been given here to problems of reliability arising from variability in performance of subjects over time—an issue not central to the present discussion

In the light of all these considerations, the conclusion seems clear that the original ratings had fully adequate reliability for the purpose for which they were used.[2]

References

Cohen, D., Whitmyre, J. W., and Funk, W. H. Effect of group cohesiveness and training upon creative thinking. (Abstract) Amer. Psychol., 1959, 14, 410-411. See also: Cohen, D., Whitmyre, J. W., and Funk, W. H. Effect of group cohesiveness and training upon creative thinking. J. appl. Psychol., 1960, 319-322.

Taylor, D. W., Berry, P. C., and Block, C. H. Does Group Participation when Using Brainstorming Facilitate or Inhibit Creative Thinking? New Haven: Depts. Industr. Admin. and Psychol., 1957 (Tech. Rep. 1, Contract Nonr 609-20). Subsequently published in: Administrative Science Quarterly, 3, 1958, 23-47.

[2] If for some purpose higher reliabilities are needed, they may be obtained, of course, employing these scales by having two or more raters rate each response and then taking the sum or mean of these ratings as the rating for that response. The reliabilities which would be expected may be estimated from the present coefficients by employing the Spearman-Brown formula.